ST5225: Statistical Analysis of Networks
Lecture 7: WWW network

WANG Wanjie
staww@nus.edu.sg

Department of Statistics and Applied Probability
National University of Singapore (NUS)

Saturday 17 March, 2018

Outline

- Review
- World Wide Web network, part I

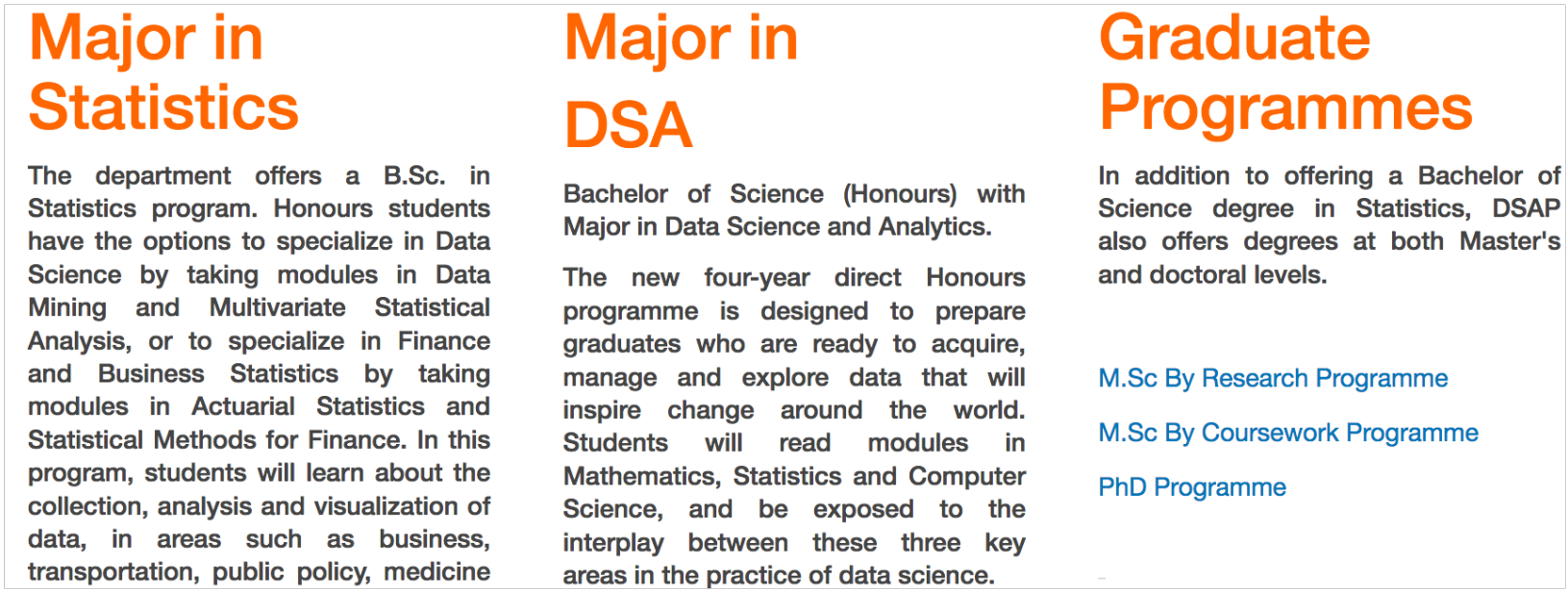
Review

- Market clearing prices
  - What is market-clearing prices
  - How to construct market clearing prices
- Bargaining on networks
  - Experiment: dividing resources on an edge (describe the process of this experiment, and the results for two-nodes network and three-nodes network)
  - Nash Bargaining Solution: outside option, split (apply the formula to problems)
  - Stable outcomes, instability (check whether an outcome is stable or not, and specify the reason)
  - Natural/Balanced stable outcomes (check whether the outcome is balanced or not)
  - Ultimatum game (know this story)

Overview

- Introduction
  - What is world wide web network (Understand the meanings of nodes/links of the network)
  - Review: strongly connected; bowtie structure (Identify the parts of bowtie structure)
- Web search
  - Authority and Hub (do 1-2 iterations by hand for small network; and run with code for large network)
  - Page Rank (do 1-2 iterations by hand for small network; and run with code for large network)
  - Mathematics (Know why we have these mathematical results)
- Advertisement
  - Have the idea what is the problem for advertisements
  - Relate the idea of advertisement price with the matching market problem
  - Construct the price for advertisements

- Nowadays, surfing on the internet is an important part of our life
- Say, when you consider applying for the M.Sc. programme in NUS DSAP, you will check the corresponding website



Major in Statistics

The department offers a B.Sc. in Statistics program. Honours students have the options to specialize in Data Science by taking modules in Data Mining and Multivariate Statistical Analysis, or to specialize in Finance and Business Statistics by taking modules in Actuarial Statistics and Statistical Methods for Finance. In this program, students will learn about the collection, analysis and visualization of data, in areas such as business, transportation, public policy, medicine

Major in DSA

Bachelor of Science (Honours) with Major in Data Science and Analytics.

The new four-year direct Honours programme is designed to prepare graduates who are ready to acquire, manage and explore data that will inspire change around the world. Students will read modules in Mathematics, Statistics and Computer Science, and be exposed to the interplay between these three key areas in the practice of data science.

Graduate Programmes

In addition to offering a Bachelor of Science degree in Statistics, DSAP also offers degrees at both Master's and doctoral levels.

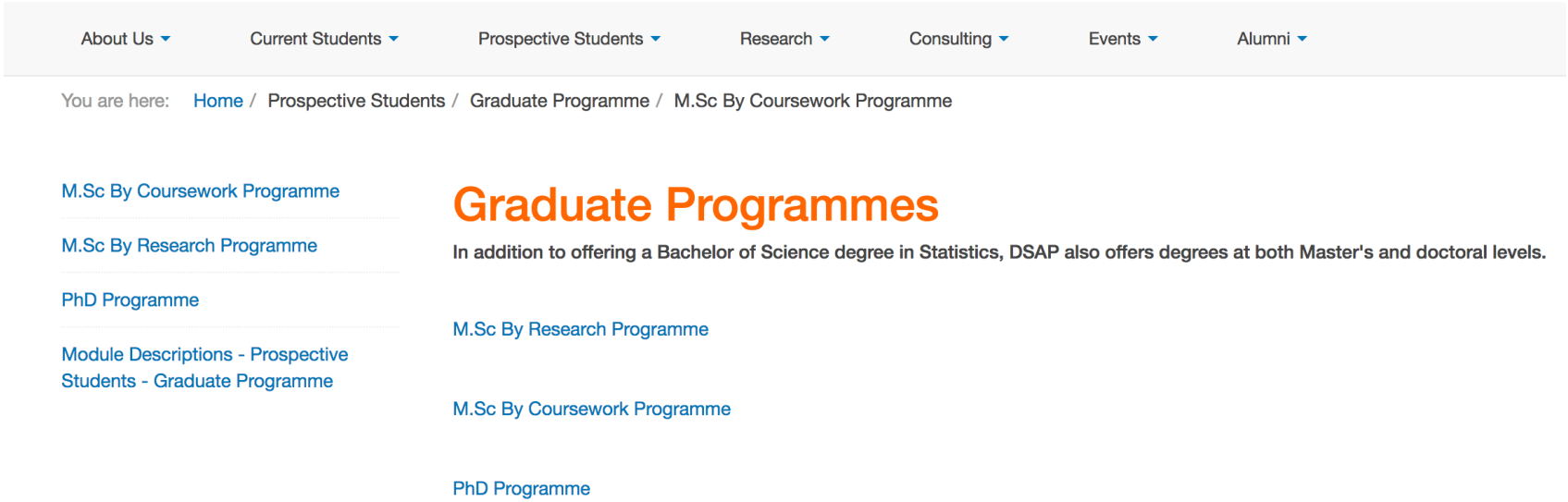
M.Sc By Research Programme

M.Sc By Coursework Programme

PhD Programme

- There is a *hyperlink* to a webpage containing the M.Sc. programme information

- Click on the programme website (part of it)



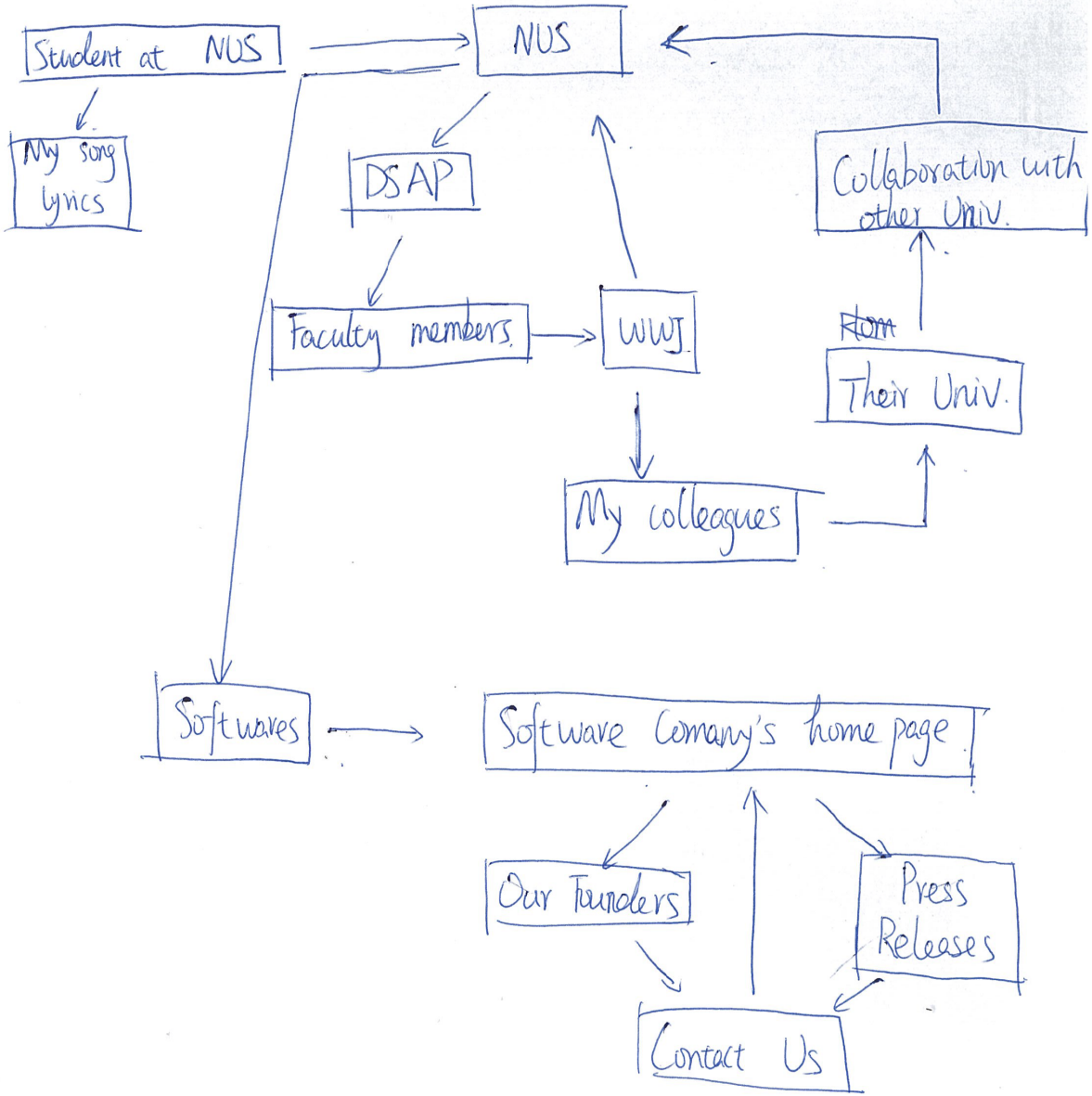
- There are a bunch more hyperlinks, directing the reader to other webpages, such as other programmes, module descriptions, etc.
- The "home" hyperlink will direct us back
- It is possible there is no hyperlinks on one webpage.
- There are also some hidden hyperlinks. If you put the mouse on the header, then more links are coming out.

World Wide Web Network

- These hyperlinks could form a network
- WWW network is a directed network, where
  - Node: a webpage
  - Edge: the hyperlink
  - Direction of the edge: If A directs us to B, then the edge points from A towards B
  - The directions show a path we can go through
- Remarks
  - There are reciprocal edges, where A points to B, and B points back to A
  - A similar network example is citation network, where paper A points to paper B if A cites B. For the citation network, the existence of reciprocal edges is much less, since usually B is earlier than A.
  - Example: political blogs data we have mentioned

World Wide Web network

Example: A small network starting with a student in our class



World Wide Web network

Recall:

- in-degree: number of edges pointing toward node A
  out-degree: number of edges starting from node A
- Adjacency matrix: asymmetric
- Strongly connected: A directed graph is called *strongly connected* if there is a directed path from A to B for any two nodes A and B in the graph
- Bowtie structure: Strongly connected component, in-component, out-component, tubes, tendrils (see the example)
- The bowtie structure is very natural in WWW network, because the main page of major commercial, governmental, non-profit organizations in the world usually have a directory connected with each other and also many other organizations.

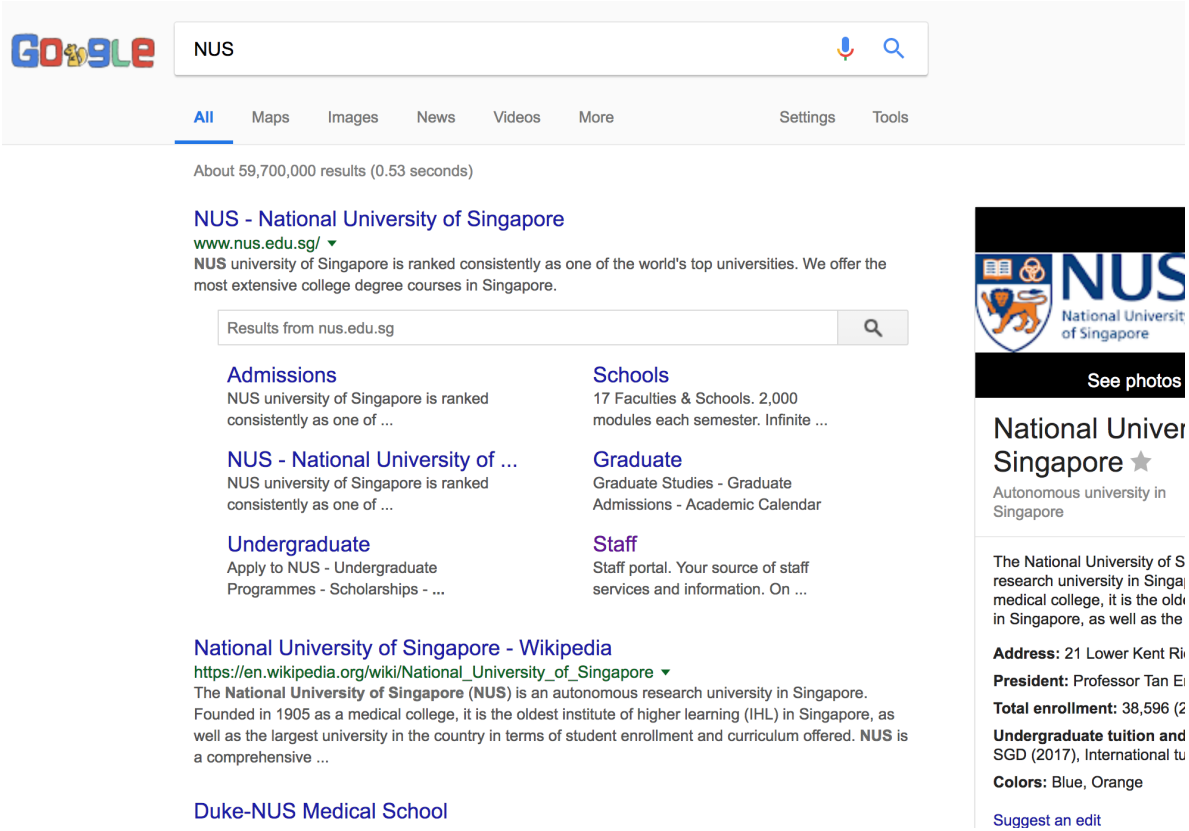
World Wide Web network

Recall:

- in-degree: number of edges pointing toward node A
  out-degree: number of edges starting from node A
- Adjacency matrix: asymmetric
- Strongly connected: A directed graph is called *strongly connected* if there is a directed path from A to B for any two nodes A and B in the graph
- Bowtie structure: Strongly connected component, in-component, out-component, tubes, tendrils (see the example)
- The bowtie structure is very natural in WWW network, because the main page of major commercial, governmental, non-profit organizations in the world usually have a directory connected with each other and also many other organizations.

Search engine: Page ranking

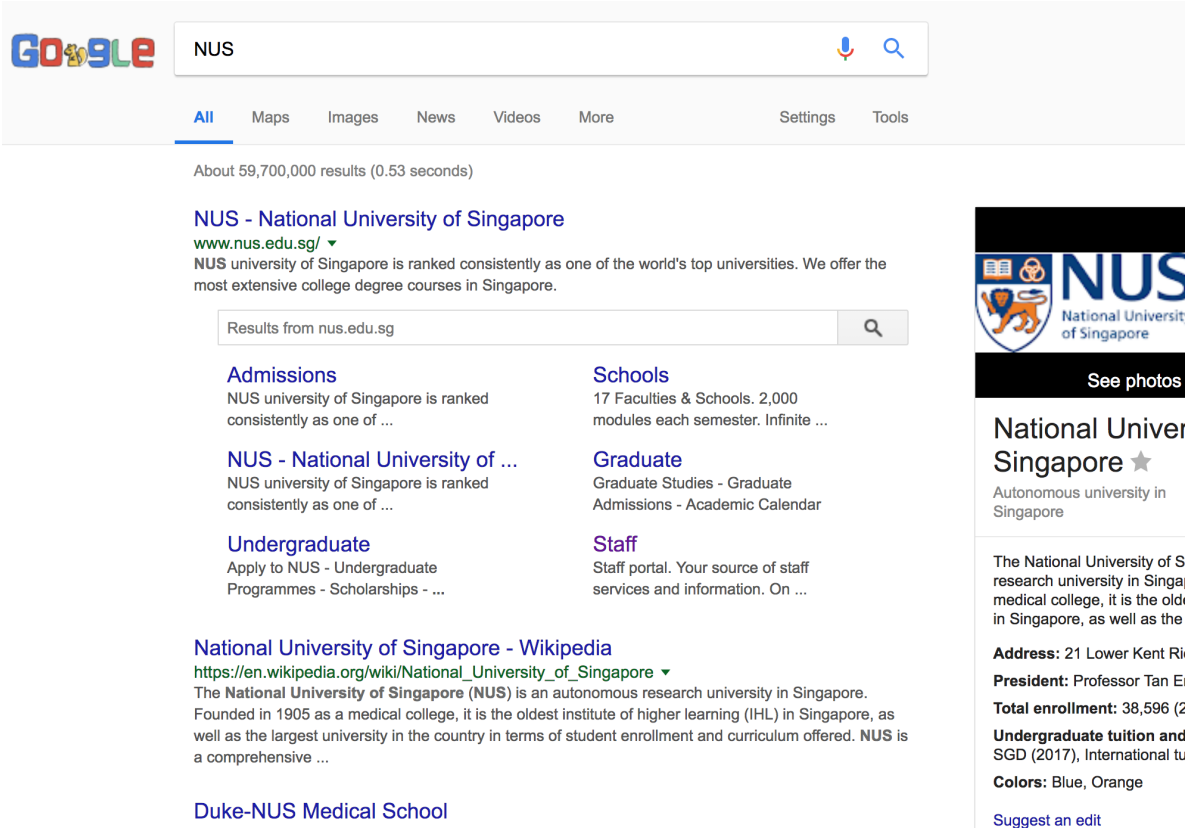
- Search "NUS" in google



- The top one is the homepage of NUS, but how google knows it?
- Obviously, it cannot be done by hand...

Search engine: Page ranking

- Search "NUS" in google

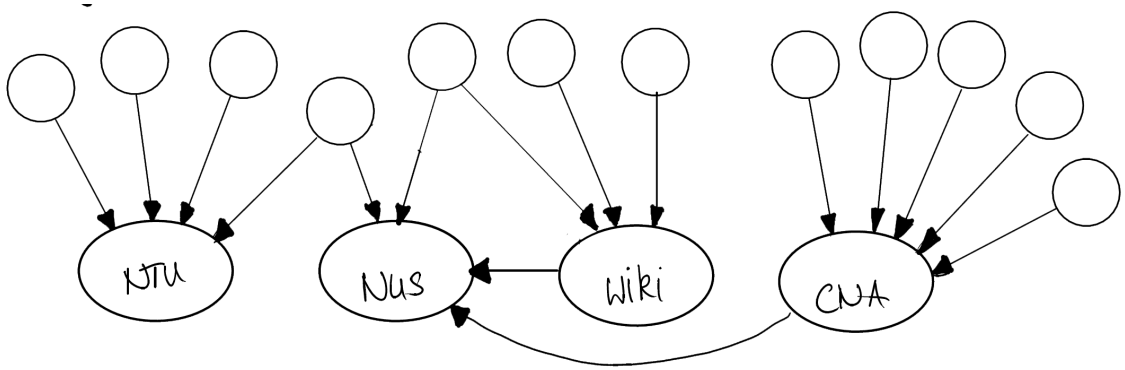


- The top one is the homepage of NUS, but how google knows it?
- Obviously, it cannot be done by hand...

Information Retrieval

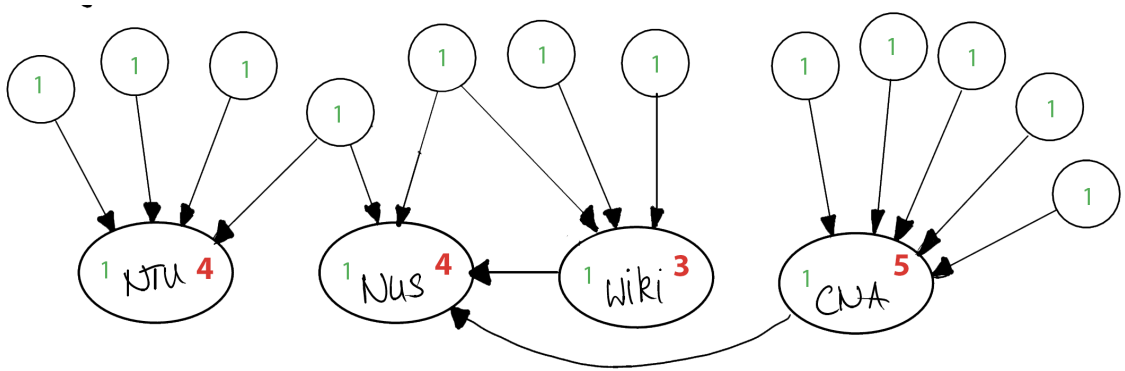
- How to search the information with some keywords? This field is called *information retrieval*, and has been studied since 1960s
- Difficulties:
  - Synonymy: aircraft/plane; thermodynamics/heat; etc.
  - Polysemy: plane (airplane/term in geometric/surface)
  - Personalization: different people may interest in different things. Searching for airplane, general people may want to check flights, technical people may be interested in the structure, etc.
  - Timing: google news
- We cannot solve all the problems. Here is just one possible solution.

Searching for NUS



- The webpage itself cannot tell us anything
  - Wiki may contain more "NUS" than the homepage of NUS
- Clicks?
  - Obviously, the homepage of nus is not the one with highest clicks.
  - If weighting with clicks, then one can easily increase the clicks for a false website
- Links are better
  - Webpages received more in-links are more important
  - Consider the in-links from the related webpages only
  - We call this as the *authority* of the webpage

Link analysis

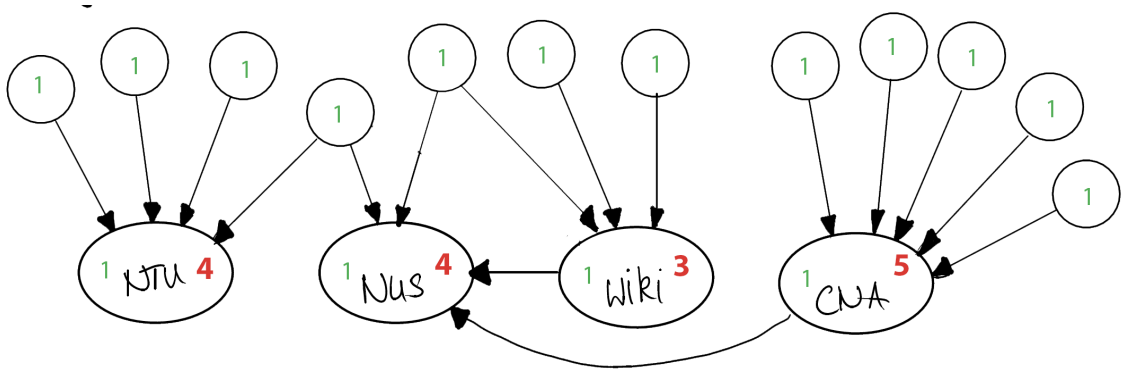


Procedure:

1. For a keyword, say, NUS, figure out all the webpages related to this word
2. For each webpage, calculate the number of in-links (in-degree)
3. For this small network, the webpage with largest in-degree is selected to be the top

Problem. With this toy example, CNA is selected since the news may be cited many times...

Link analysis



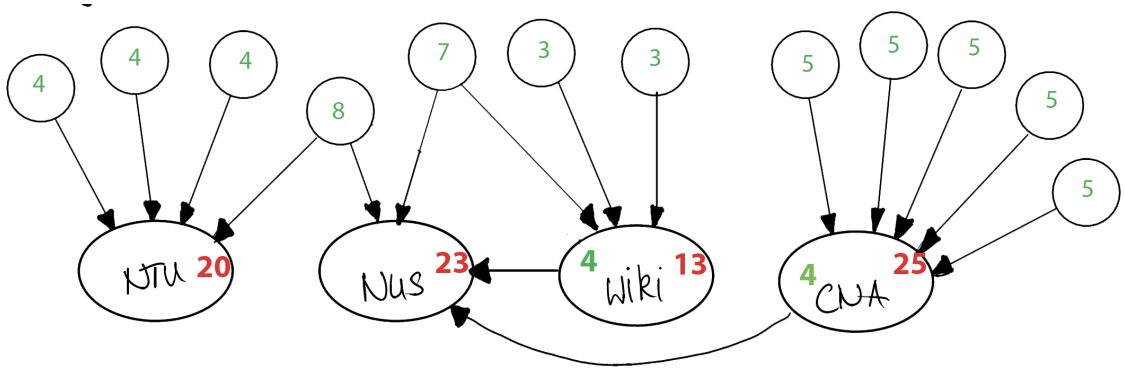
Procedure:

1. For a keyword, say, NUS, figure out all the webpages related to this word
2. For each webpage, calculate the number of in-links (in-degree)
3. For this small network, the webpage with largest in-degree is selected to be the top

Problem. With this toy example, CNA is selected since the news may be cited many times...

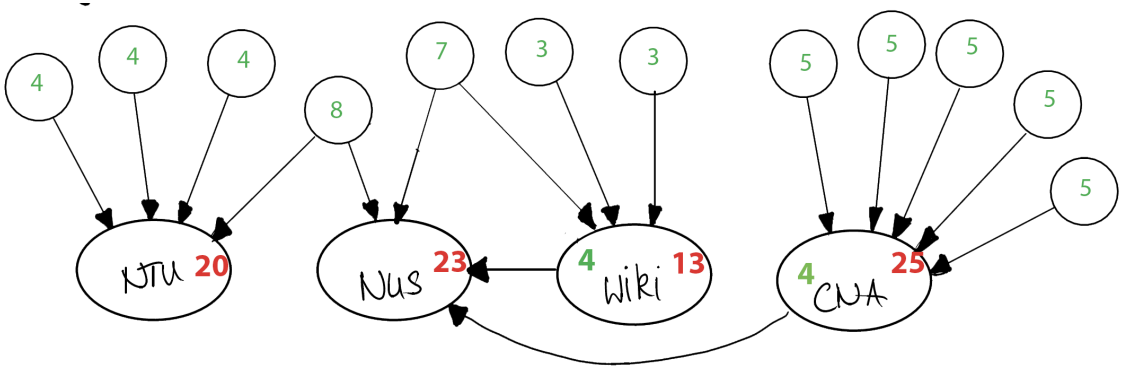
Link analysis

- Solution. We also evaluate the importance of each webpage



- Webpages that listed more related websites should have larger weights
- Evaluate the importance of the webpages according to their out-links (out-degrees) to the authorised webpages
- With the updated evaluation, calculate the authority for each webpage again
- This is called the *hub* of the webpages

Link analysis

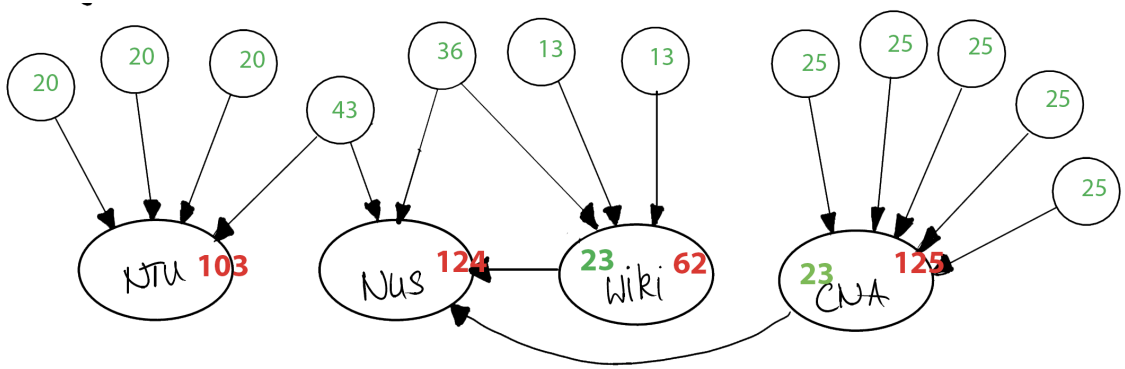


Procedure.

1. Each webpage begin with authority 1 and hub 1
2. Calculate the authority of each webpage
3. Calculate the hub for each webpage, to be the sum of authorities of all pages it points to
4. Update the authority of each page, to be the sum of hubs of all pages that point to it
5. The updated authority gives the rank of webpages

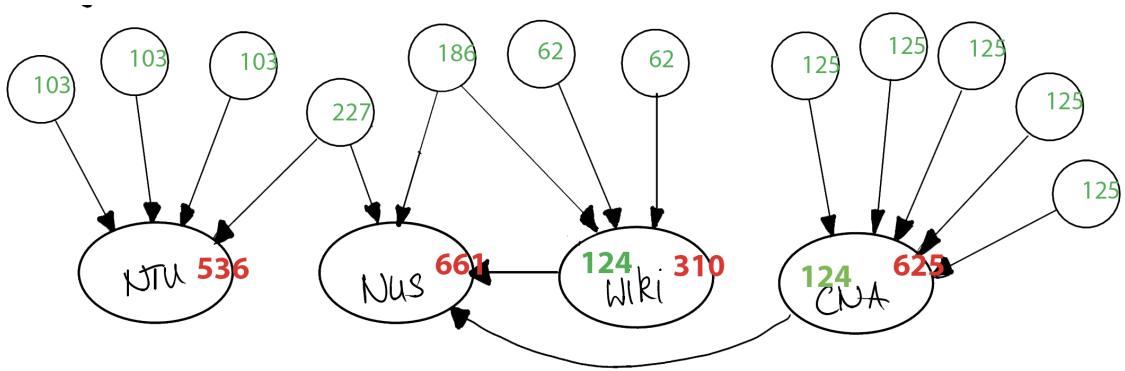
The importance of NUS increases, but still not as large as CNA

Repeat the procedure again and again.



1. Each webpage begin with authority 1 and hub 1
2. **Authority Update Rule**: For each webpage, update the authority to be the sum of hubs of all pages that point to it
3. **Hub Update Rule**: For each webpage, update the hub to be the sum of authorities of all pages that it points to
4. Repeat the Authority Update Rule first, Hub Update Rule second updates for k times, where k is pre-selected
5. The updated authority gives the rank of webpages

Repeat the procedure again and again.



- When $k = 3$, the homepage of NUS becomes the top one
- Numbers are too big? Normalize!
- Note that hubs and authorities denote different properties of the webpages
  - Authority: how many other resources recognize this webpage as important
  - Hub: Reliability of the webpage citing others. Is it a hub that cite many webpages with the same keyword?
  - The hub score is higher if it cited high-authority pages

Authority-Hub-algorithm.

1. Each page v has two scores, $auth(v)$, $hub(v)$
2. Start with $hub(v) = 1$ for each v
3. Repeat
   - Normalize $hub(v)$ so that $\sum_v hub(v) = 1$
   - For each v, update $auth(v) = \sum_{u,(u,v)\in E} hub(u)$
   - Normalize $auth(v)$ so that $\sum_v auth(v) = 1$
   - For each v, update $hub(v) = \sum_{u,(v,u)\in E} auth(u)$
4. Output the result according to the authorities

Link analysis: Mathematics

- This idea sounds familiar: evaluate a node according to its neighbors
- Eigenvector centrality for undirected graph
- Directed graph?

For directed graphs, there is also an adjacency matrix A, where

$$A_{ij} = \begin{cases} 1, & \text{if } (i,j) \in E \\ 0, & \text{otherwise} \end{cases}$$

- $A_{ij} = 1$ means there is an edge from i to j
- A is not symmetric for directed networks

Link analysis: Mathematics

- Let H denote the hub vector, where $H(j) = Hub(j)$; let $Auth$ denote the authority vector. Both vectors have length $|V|$.
- Hub Update Rule

$$H(v) = \sum_{u,(v,u)\in E} auth(u) = \sum_{u,A_{vu}=1} auth(u) = \sum_{u} A_{vu} Auth(u)$$

  Therefore, $H = A \times Auth$
- Authority Update Rule

$$Auth(v) = \sum_{u,(u,v)\in E} Hub(u) = \sum_{u,A_{uv}=1} Hub(u) = \sum_{u} H(u) A_{uv}$$

  Therefore, $Auth = A^T \times H$
- For a repetition,

$$H^{(new)} = A \times Auth = A \times A^T \times H = AA^T H,$$
$$Auth^{(new)} = A^T \times H = A^T \times A \times Auth = A^T A \times Auth$$

Link analysis: Mathematics

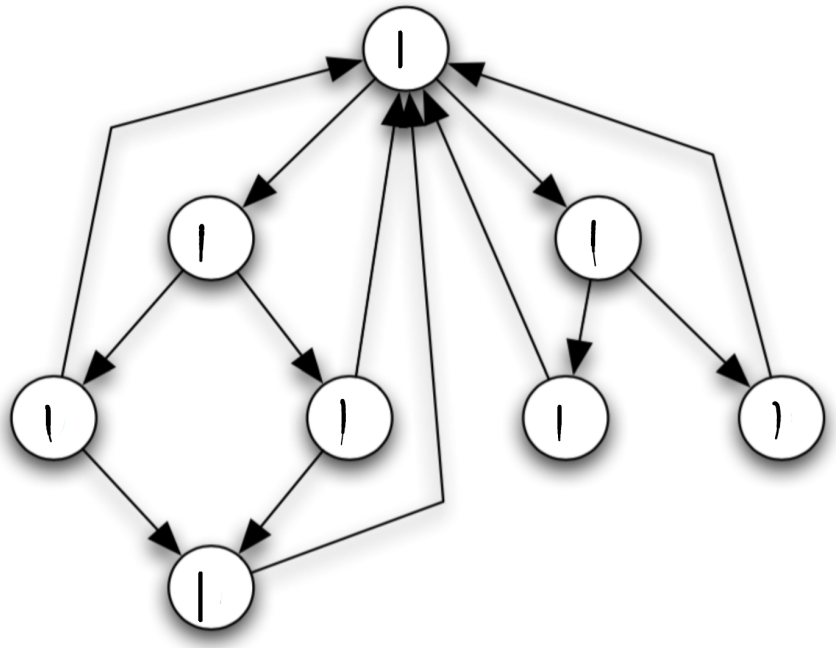
- Begin with $H = (1/|V|, 1/|V|, \cdots, 1/|V|)^T$
- After k steps, $H^{(k)} = (AA^T)^k H$
- $Auth^{(0)} = A^T \times H$ and then normalize. After k steps, $Auth^{(k)} = (A^T A)^k Auth^{(0)}$

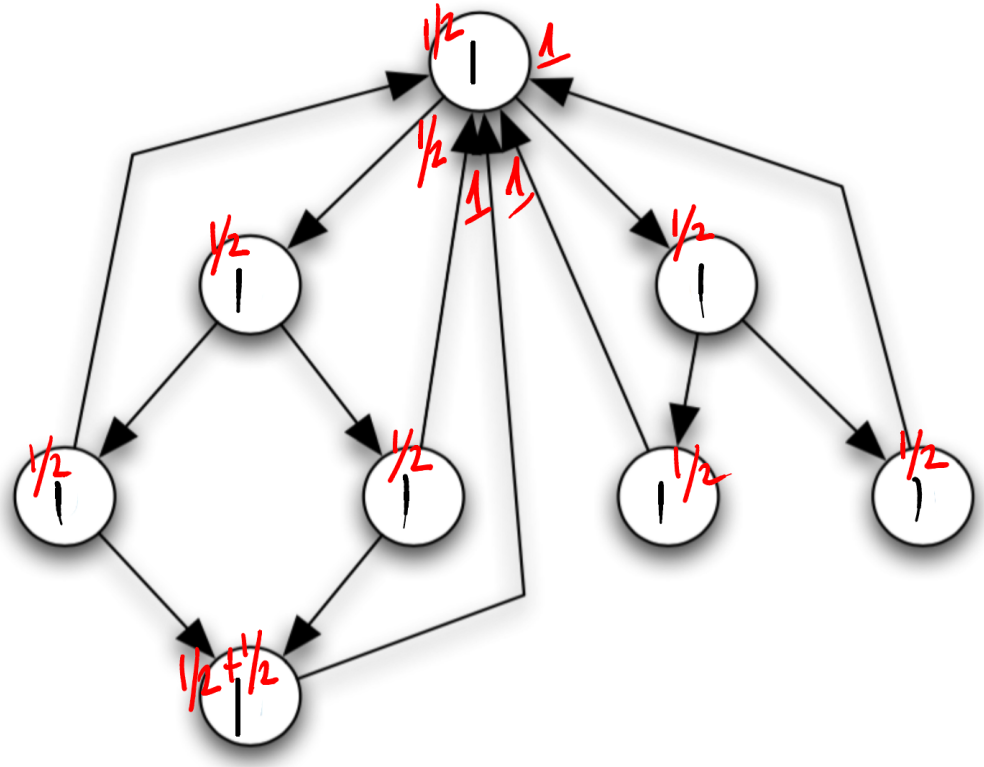
- When $k \to \infty$, H converges to the top eigenvector of AA^T, and $Auth$ converges to the top eigenvector of $A^T A$.
- Note: these two eigenvectors are not the same!!

- In short, we may calculate the eigenvector of AA^T or $A^T A$ directly, and gives the ranks for pages

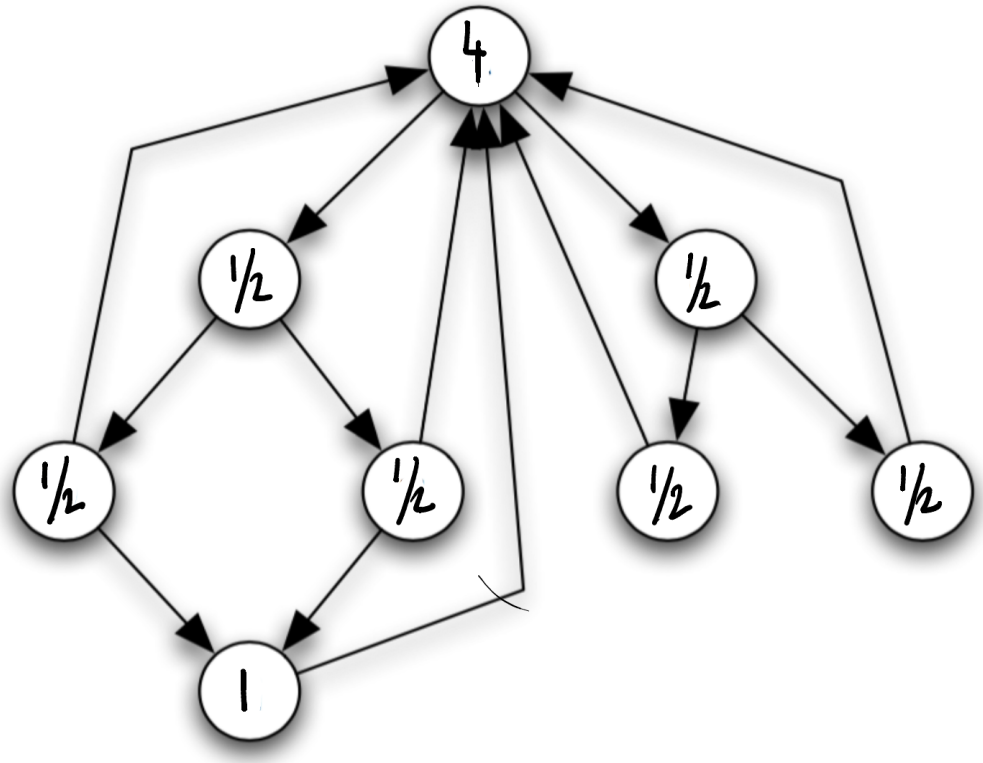
- We consider the hubs and authorities, different properties related to the same webpage
- Can we combine these two and consider the "importance" only?
- Solution: PageRank (Google)

- Idea: Assume there is a fluid in the network, flows from nodes to nodes, and we take the nodes with most fluid
- Step 1. Each page starts with PageRank of 1.





- Step 2. Each page pass the scores to the other pages. For each node v, if the PageRank is $p(v)$ and the out-degree is K, then it passes $p(v)/K$ to each neighbour.



- Step 3. Update the PageRank of each node according to the scores received
- Step 4. Repeat the "passing-receiving-updating" procedure for k steps

- Usually, at the start point, the PageRank is assigned to be with sum 1. In the previous example, it is 1/8 for each node. For better illustration, I use 1 for each node.
- The weights for nodes/edges at the start point can be manually assigned to be unequal
- Note that there is no loss or gain of PageRank. The fluid flows from this node to that node. So theoretically, no normalization is required
- However, in reality, a "scaled" version of PageRank is applied:
  1. After each repetition, scale the PageRank to be $p(v) \times s$
  2. Add $\frac{1-s}{|V|}$ to each node
- In practice, $s \sim 0.9$
- The scaled PageRank could avoid absorbing strongly connected components

PageRank: Random Walk

Another way to understand the PageRank.

- Assuming a student Alice is browsing a network of webpages
- First, Alice chooses a page at random
- Then, she randomly picks a link, and follows the link to the next page, with equal probability
- Alice re-do the procedure for k steps

- The PageRank gives the probability that Alice stays at a given webpage after k steps

PageRank: Mathematics

- Recall that in stochastic process, we use the transition matrix to calculate the probability the random walk stays at each state
- Here, we have the transition matrix as

$$P_{ij} = \begin{cases} \frac{1}{d_i^{out}} = \frac{1}{\sum_k A_{ik}}, & \text{if } (i,j) \in E \\ 0, & \text{otherwise} \end{cases}$$

- At the starting point, the distribution on the nodes is

$$r = (\frac{1}{|V|}, \frac{1}{|V|}, \cdots \frac{1}{|V|})$$

- After k steps, the distribution on the nodes is

$$r^{(k)} = rP^k$$

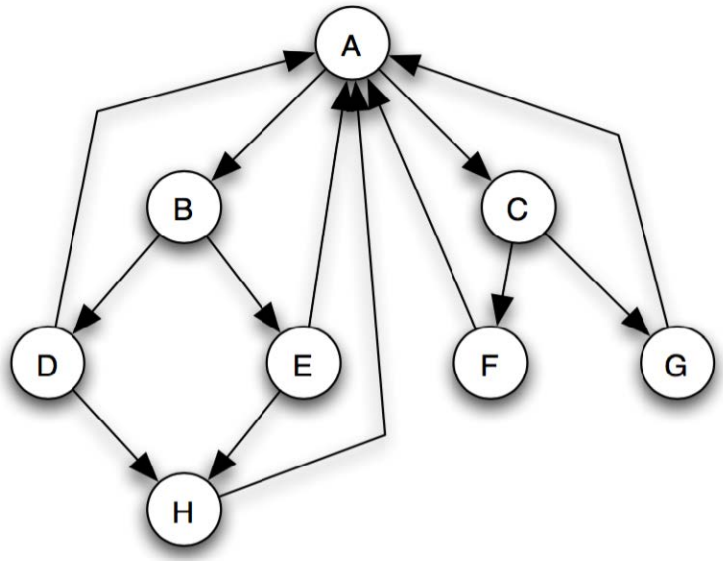
- According to the theory in stochastic processes, if the graph satisfies some conditions (positive recurrent), then there is a limiting distribution when $k \to \infty$, where

$$\pi = \lim_{k \to \infty} rP^k, \qquad \pi = \pi P$$

PageRank: Random Walk

Adjacency matrix

$$A = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \\ G \\ H \end{array} \begin{array}{c} \begin{array}{cccccccc} A & B & C & D & E & F & G & H \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}$$



PageRank: Random Walk

- out-degrees: $d_A = 2$, $d_B = 2$, $d_C = 2$, $d_D = 2$, $d_E = 2$, $d_F = 1$, $d_G = 1$, $d_H = 1$,
- Transition matrix:

$$P = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \\ F \\ G \\ H \end{array} \begin{array}{c} \begin{array}{cccccccc} A & B & C & D & E & F & G & H \end{array} \\ \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/2 & 1/2 & 0 \\ 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 \\ 1/2 & 0 & 0 & 0 & 0 & 0 & 0 & 1/2 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

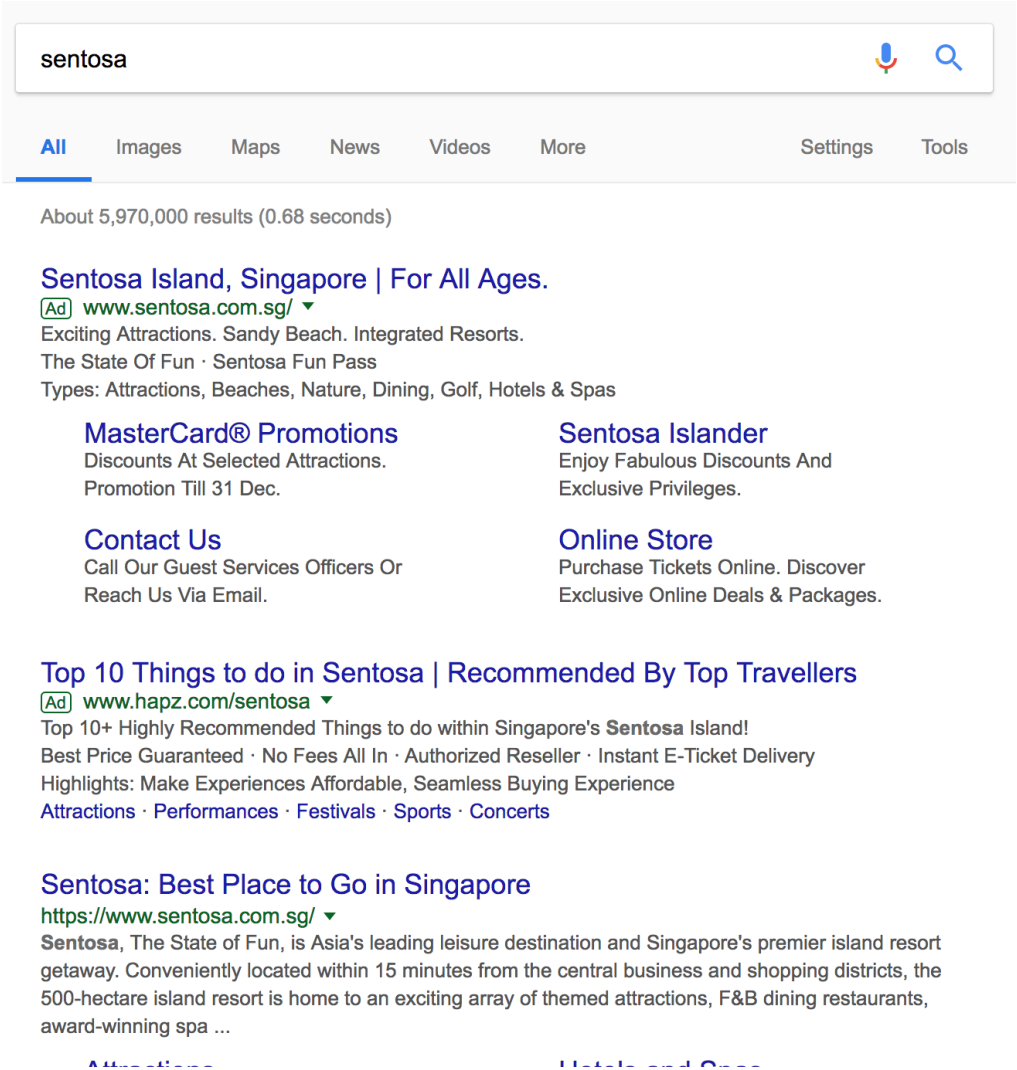
- Starting probability: $r = (1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8, 1/8)$
- Probability after 1 step:
  $rP = (1/2, 1/16, 1/16, 1/16, 1/16, 1/16, 1/16, 1/8)$
- Limiting dist: $\pi = (0.30, 0.15, 0.15, 0.08, 0.08, 0.08, 0.08, 0.09)$

Summary

- Authority-Hub-Algorithm: assume there are two properties related to one webpage, the authority (reliability when introduce this topic) and the hub (reliability when review this topic). Update the authority score and hub score each time
- PageRank: assume there is only one importance score related to each webpage, and the importance score (PageRank) is decided by a random walk on the network
- Scaled PageRank: to avoid being trapped by the absorbing components, re-scale the scores every time
- Mathematical expression for the methods
- Generalised to the popular fields nowadays, say, find the hot topics on Weibo/Twitter, etc.

Advertisement

- If you search "Sentosa" on google, the results will have two advertisement, followed by the results



Advertisement

- The companies, say Sentosa trip agents, pay Google to post an advertisement about them
- When does google display these ads?
- People who search for "Sentosa" would be more interested in the Sentosa trips
- When people search for "Sentosa", Google displays the ads according to the money paid
- People who are interested in the ads may click to the link

- It happens for every search engine.
- Without the search engines, companies pay to display ads for everyone. Now they show ads to those who have intent

Advertisement

- The companies, say Sentosa trip agents, pay Google to post an advertisement about them
- When does google display these ads?
- People who search for "Sentosa" would be more interested in the Sentosa trips
- When people search for "Sentosa", Google displays the ads according to the money paid
- People who are interested in the ads may click to the link

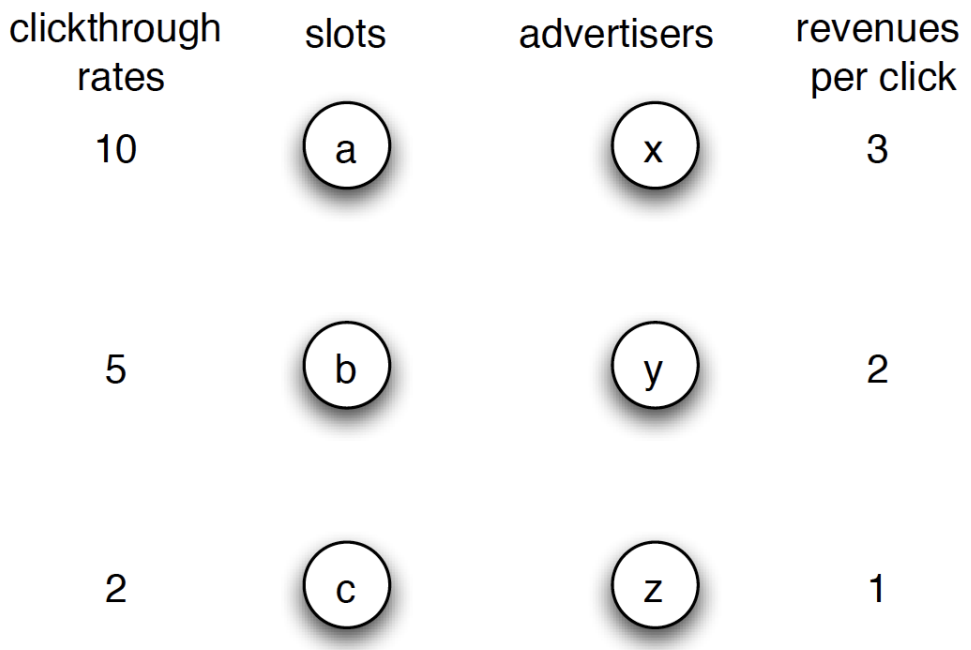
- It happens for every search engine.
- Without the search engines, companies pay to display ads for everyone. Now they show ads to those who have intent

Advertisement

Settings and Problems:

- The company creates an ad that shows every time when the user enters "Sentosa", and that ad links to their website. *The company pay the search engine only when some one clicks through the link.* This strategy is called *paying per click*.
  - Examples in the textbook:
  - "Calligraphy pens": \$1.70 per click
  - "mortgage refinancing": \$50 per click
- How to set the price for each company?
  - Too many keywords, and it is hard to set price for each search
  - Auction the slots would help, yet how to set the auction?
  - Vickrey-Clarke-Groves (VCG) mechanism

Matching Market



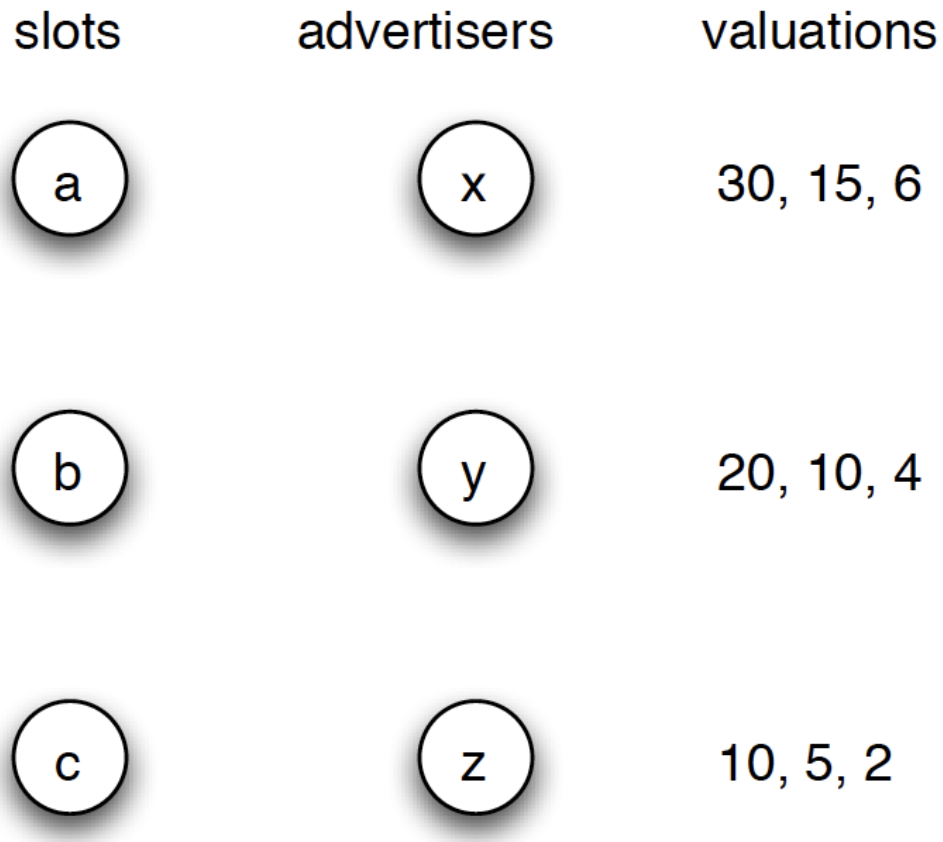
- Several slots for one keyword. Top slots indicate high clichthrough rates.
- Several advertisers. Each advertiser gets different revenue from clicks
- **Target**: Charge on slots properly

Matching Market

- For one advertiser, the valuation of a slot is

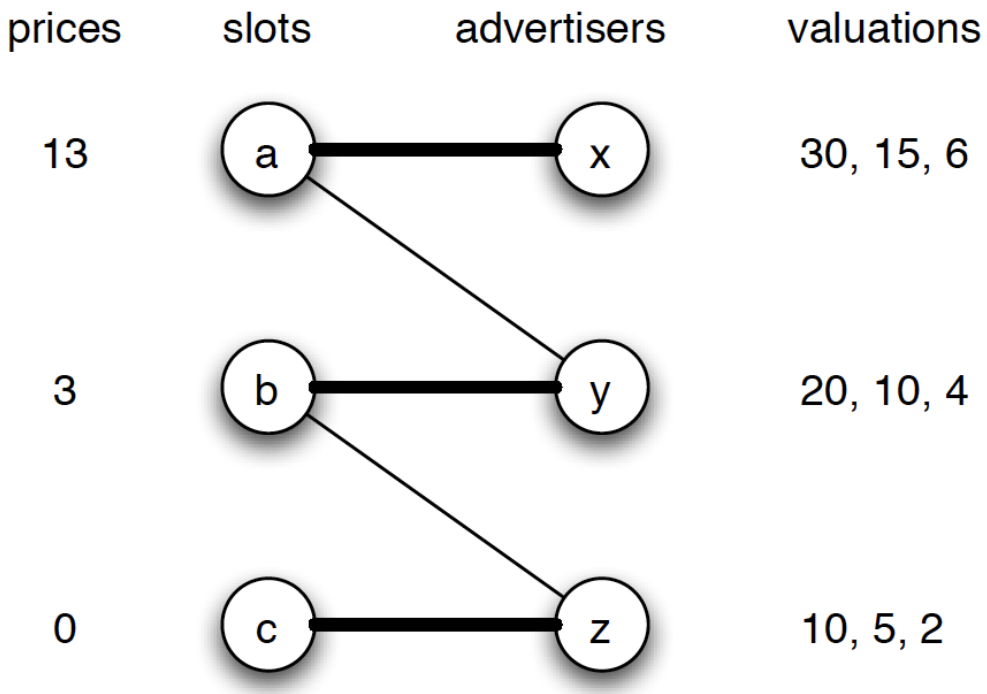
$$\text{valuation} = \text{clickthrough rate} \times \text{revenue per click}$$

- Matching market problem!



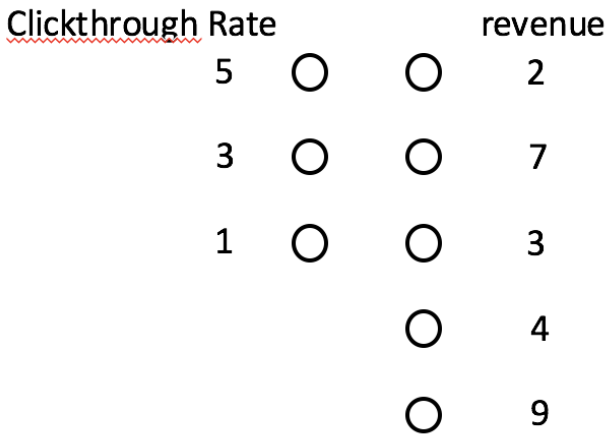
Matching Market

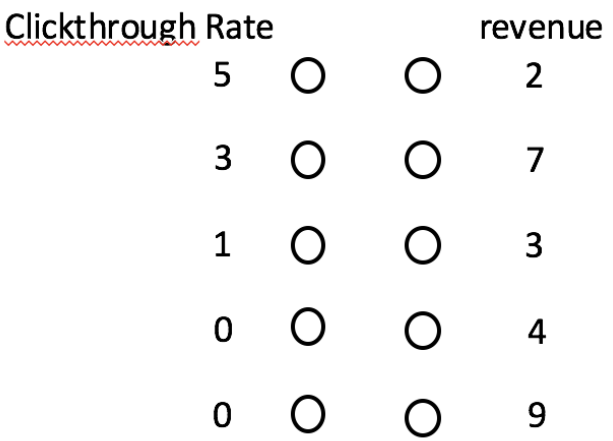
- Find a series of market-clearing prices, and the corresponding matching.



Matching Market

- More slots than advertises, or more advertisers than slots?
- Create "fake" slots and "fake" advertisers.
  - Advertisers $>$ slots: create "fake" slots of clickthrough rate 0
  - Advertisers $<$ slots: create "fake" advertisers of revenue 0.





Summary

- The advertisement problem for WWW network
- How to formulate the problem
- Solution: matching markets
- In reality, the advertisers may lie on the revenue. One solution is to apply VCG mechanism. Check the textbook if you are interested.
- More topics of interest. e.g., the "hidden" clickthrough rates.